

Learning from Multi-View Structural Data via Structural Factorization Machines

Chun-Ta Lu

University of Illinois at Chicago
Chicago, IL
clu29@uic.edu

Hao Ding

Beijing University of Posts and Telecommunications
Beijing, China
haoding.tourist@gmail.com

Lifang He

Shenzhen University
Shenzhen, China
lifanghescut@gmail.com

Philip S. Yu

University of Illinois at Chicago
Chicago, IL
psyu@cs.uic.edu

ABSTRACT

Real-world relations among entities can often be observed and determined by different perspectives/views. For example, the decision made by a user on whether to adopt an item relies on multiple aspects such as the contextual information of the decision, the item's attributes, the user's profile and the reviews given by other users. Different views may exhibit multi-way interactions among entities and provide complementary information. In this paper, we introduce a multi-tensor-based approach that can preserve the underlying structure of multi-view data in a generic predictive model. Specifically, we propose structural factorization machines (SFMs) that learn the common latent spaces shared by multi-view tensors and automatically adjust the importance of each view in the predictive model. Furthermore, the complexity of SFMs is linear in the number of parameters, which make SFMs suitable to large-scale problems. Extensive experiments on real-world datasets demonstrate that the proposed SFMs outperform several state-of-the-art methods in terms of prediction accuracy and computational cost.

CCS CONCEPTS

•Information systems →Data mining; •Computing methodologies →Machine learning;

KEYWORDS

Tensor Factorization; Structural Data; Multi-View Learning

ACM Reference format:

Chun-Ta Lu, Lifang He, Hao Ding, and Philip S. Yu. 2017. Learning from Multi-View Structural Data via Structural Factorization Machines. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17)*, 9 pages.

DOI: 10.1145/nnnnnnn.nnnnnnn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM...\$15.00

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the ability to access massive amounts of heterogeneous data from multiple sources, multi-view data have become prevalent in many real-world applications. For instance, in recommender systems, online review sites (like Amazon and Yelp) have access to contextual information of shopping histories of users, the reviews written by the users, the categorizations of the items, as well as the friends of the users. Each view may exhibit pairwise interactions (e.g., the friendships between users) or even higher-order interactions (e.g., a customer write a review for a product) among entities (such as customers, products, and reviews), and can be represented in a multi-way data structure, i.e., tensor. Since different views usually provide complementary information [4, 5, 17], how to effectively incorporate information from multiple structural views is critical to good prediction performance for various machine learning tasks.

Typically, a predictive model is defined as a function of predictor variables (e.g., the customer id, the product id, and the categories of the product) to some target (e.g., the rating). The most common approach in predictive modeling for multi-view structural data is to describe samples with feature vectors that are flattened and concatenated from structural views, and apply a machine learning method, such as linear regression (LR) and support vector machines (SVMs), to learn the target function from observed samples. Recent works have shown that linear models fail for tasks with very sparse data [21, 22]. A variety of methods have been proposed to address the data sparsity issue by factorizing the monomials (or feature interactions) with kernel tricks, such as the ANOVA kernels used in FMs [2, 21] and polynomial kernels used in polynomial networks [3, 16]. However, the disadvantages of this approach are that (1) the important structural information of each view will be discarded which may lead to the degraded prediction performance and (2) the feature vectors can grow very large which can make learning and prediction very slow or even infeasible, especially if each view involves relations of high cardinality. For example, including the relation “friends of a user” in the feature vector (represented by their IDs) can result in a very long feature vector. Further, it will repeatedly appear in many samples that involve the given user.

Matrix/tensor factorization models have been a topic of interest in the areas of structural data analysis, such as community detection [10], collaborative filtering [13, 24], and neuroimage analysis [9]. Assuming multi-view data have the same underlying low-rank structure (at least in one mode), coupled data analysis such as collective matrix factorization (CMF) [25] and coupled matrix and tensor factorization (CMTF) [1] that jointly factorize multiple matrices (or tensors) has been applied to applications such as clustering and missing data recovery. However, since existing coupled factorization models are unsupervised, the importance of each structural view in modeling the target value cannot be automatically learned. Furthermore, when applying these models to data with rich meta information (e.g., friendships) but extremely sparse target values (e.g., ratings), it is very likely the learning process will be dominated by the meta information without manual tuning some hyperparameters, e.g., the weights of the fitting error of each matrix/tensor in the objective function [25], the weights of different types of latent factors in the predictive models [14], or the regularization hyperparameters of latent factor alignment [18].

In this paper, we propose a general and flexible framework for learning the predictive structure from the complex relationships within the multi-view structural data. Each view of an instance in this framework is represented by a tensor that describes the multi-way interactions of subsets of entities, and different views have some entities in common. Constructing the tensors for each instance may not be realistic for real-world applications in terms of space and computational complexity, and the model parameters can have exponential growth and tend to be overfitting. In order to preserve the structural information of multi-view data without physically constructing the tensors, we introduce structural factorization machines (SFMs) that can learn the consistent representations in the latent feature spaces shared in the multi-view tensors while automatically adjust the contribution of each view in the predictive model. Furthermore, we provide an efficient method to avoid redundant computing on repeating patterns stemming from the relational structure of the data, such that SFMs can make the same predictions but with largely speed up computation.

The contributions of this paper are summarized as follows:

- We introduce a novel multi-tensor framework for mining data from heterogeneous domains, which can explore the high order correlations underlying multi-view structural data in a generic predictive model.
- We develop structural factorization machines (SFMs) tailored for learning the common latent spaces shared in multi-view tensors and automatically adjusting the importance of each view in the predictive model. The complexity of SFMs is linear in the number of features, which makes SFMs suitable to large-scale problems.
- Extensive experiments on eight real-world datasets are performed along with comparisons to existing state-of-the-art factorization models to demonstrate its advantages.

The rest of this paper is organized as follows. We introduce the problem definition and some preliminary works in Section 2. We then propose the framework for learning multi-view structural data, and develop the structural factorization machines (SFMs), and

Table 1: List of basic symbols.

Symbol	Definition and description
x	each lowercase letter represents a scalar
\mathbf{x}	each boldface lowercase letter represents a vector
\mathbf{X}	each boldface uppercase letter represents a matrix
\mathcal{X}	each calligraphic letter represents a tensor
\mathfrak{X}	each gothic letter represent a general set or space
$[1 : N]$	a set of integers in the range of 1 to N inclusively.
$\langle \cdot, \cdot \rangle$	denotes inner product
\circ	denotes tensor product (outer product)
$*$	denotes Hadamard (element-wise) product

provide an efficient computing method in Section 3. The experimental results and parameter analysis are reported in Section 4. Section 5 concludes this paper.

2 PRELIMINARIES

In this section, we begin with a brief introduction to some related concepts and notation in tensor algebra, and then proceed to formulate the problem we are concerned with multi-view learning.

2.1 Tensor Basics and Notation

Tensor is a mathematical representation of a multi-way array. The order of a tensor is the number of modes (or ways). A zero-order tensor is a scalar, a first-order tensor is a vector, a second-order tensor is a matrix and a tensor of order three or higher is called a higher-order tensor. To facilitate the distinction between scalars, vectors, matrices and higher-order tensors, the type of a quantity will be reflected by its representation: scalars are denoted by lowercase letters, vectors by boldfaced lowercase letters, matrices by boldface uppercase letters, and tensors are denoted by calligraphic letters. An element of a vector \mathbf{x} , a matrix \mathbf{X} , or a tensor \mathcal{X} is denoted by x_i , $x_{i,j}$, $x_{i,j,k}$, etc., depending on the number of modes. All vectors are column vectors unless otherwise specified. For an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$, its i -th row and j -th column vector are denoted by \mathbf{x}^i and \mathbf{x}_j , respectively. Given two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{I \times J}$, $\mathbf{X} * \mathbf{Y}$ denotes the element-wise (Hadamard) product between \mathbf{X} and \mathbf{Y} , defined as the matrix in $\mathbb{R}^{I \times J}$. For each integer $M > 1$, $[1 : M]$ denotes all integers from 1 to M . An overview of the basic symbols used in this paper can be found in Table 1.

Definition 2.1 (Inner product). The inner product of two same-sized tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_M}$ is defined as the sum of the products of their entries:

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \dots \sum_{i_M=1}^{I_M} x_{i_1, i_2, \dots, i_M} y_{i_1, i_2, \dots, i_M}. \quad (1)$$

Definition 2.2 (Outer product). The outer product of two tensors $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ and $\mathcal{Y} \in \mathbb{R}^{I'_1 \times I'_2 \times \dots \times I'_M}$ is a $(N + M)$ -th-order tensor denoted by $\mathcal{X} \circ \mathcal{Y}$, and the elements are defined by

$$(\mathcal{X} \circ \mathcal{Y})_{i_1, i_2, \dots, i_N, i'_1, i'_2, \dots, i'_M} = x_{i_1, i_2, \dots, i_N} y_{i'_1, i'_2, \dots, i'_M} \quad (2)$$

for all values of the indices.

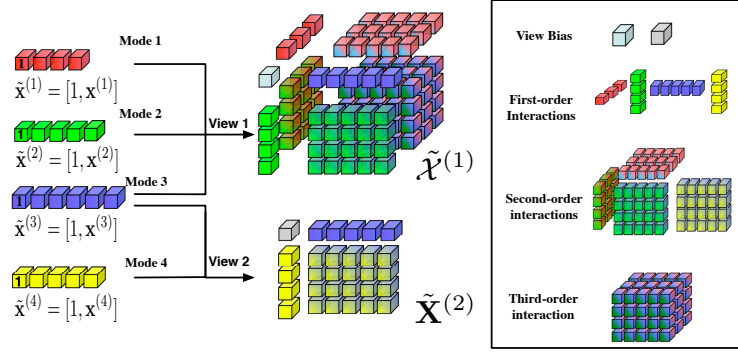


Figure 1: Example of multiple structural views, where $\tilde{\mathcal{X}}^{(1)} = \tilde{\mathbf{x}}^{(1)} \circ \tilde{\mathbf{x}}^{(2)} \circ \tilde{\mathbf{x}}^{(3)}$ and $\tilde{\mathcal{X}}^{(2)} = \tilde{\mathbf{x}}^{(3)} \circ \tilde{\mathbf{x}}^{(4)}$.

Notice that for rank-one tensors $\mathcal{X} = \mathbf{x}^{(1)} \circ \mathbf{x}^{(2)} \circ \dots \circ \mathbf{x}^{(M)}$ and $\mathcal{Y} = \mathbf{y}^{(1)} \circ \mathbf{y}^{(2)} \circ \dots \circ \mathbf{y}^{(M)}$, it holds that

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \mathbf{x}^{(1)}, \mathbf{y}^{(1)} \rangle \langle \mathbf{x}^{(2)}, \mathbf{y}^{(2)} \rangle \dots \langle \mathbf{x}^{(M)}, \mathbf{y}^{(M)} \rangle. \quad (3)$$

Definition 2.3 (CP factorization [12]). Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_M}$ and an integer R , the CP factorization is defined by factor matrices $\mathbf{X}^{(m)} \in \mathbb{R}^{I_m \times R}$ for $m \in [1 : M]$, respectively, such that

$$\mathcal{X} = \sum_{r=1}^R \mathbf{x}_r^{(1)} \circ \mathbf{x}_r^{(2)} \circ \dots \circ \mathbf{x}_r^{(M)} = \llbracket \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(M)} \rrbracket, \quad (4)$$

where $\mathbf{x}_r^{(m)} \in \mathbb{R}^{I_m}$ is the r -th column of the factor matrix $\mathbf{X}^{(m)}$, and $\llbracket \cdot \rrbracket$ is used for shorthand notation of the sum of rank-one tensors.

2.2 Problem Formulation

Our problem is different from conventional multi-view learning approaches where multiple views of data are assumed independent and disjoint, and each view is described by a vector. We formulate the multi-view learning problem using coupled analysis of multi-view features in the form of multiple tensors.

Suppose that the problem includes V views where each view consists of a collection of subsets of entities (such as person, company, location, product) and different views have some entities in common. We denote a view as a tuple $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(M)})$, $M \geq 2$, where $\mathbf{x}^{(m)} \in \mathbb{R}^{I_m}$ is a feature vector associated with the entity m . Inspired by [5], we construct tensor representation for each view over its entities by

$$\tilde{\mathcal{X}} = \tilde{\mathbf{x}}^{(1)} \circ \tilde{\mathbf{x}}^{(2)} \circ \dots \circ \tilde{\mathbf{x}}^{(M)} \in \mathbb{R}^{(1+I_1) \times \dots \times (1+I_M)},$$

where $\tilde{\mathbf{x}}^{(m)} = [1; \mathbf{x}^{(m)}] \in \mathbb{R}^{1+I_m}$ and \circ is the outer product operator. In this manner, the full-order interactions¹ between entities are embedded within the tensor structure, which not only provides a unified and compact representation for each view, but also facilitate efficient design methods. Fig. 1 shows an example of two structural views, where the first view consists of the full-order interactions among the first three modes (e.g., review text, item ID, and user ID), and the second view consists of the full-order interactions among the last two modes (e.g., user ID and friend IDs).

¹Full-order interactions range from the first-order interactions (i.e., contributions of single entity features) to the highest-order interactions (i.e., contributions of the outer product of features from all entities).

After generating the tensor representation for each view, we define the multi-view learning problem as follows. Given a training set $\mathcal{D} = \{(\{\tilde{\mathcal{X}}_n^{(1)}, \tilde{\mathcal{X}}_n^{(2)}, \dots, \tilde{\mathcal{X}}_n^{(V)}\}, y_n) \mid n \in [1 : N]\}$, where $\tilde{\mathcal{X}}_n^{(v)} \in \mathbb{R}^{(1+I_1) \times \dots \times (1+I_{M_v})}$ is the tensor representation in the v -th view for the n -th instance, y_n is the response of the n -th instance, M_v is the number of the constitutive modes in the v -th view, and N is the number of labeled instances. We assume different views have common entities, thus the resulting tensors will share common modes, e.g., the third mode in Fig 1. As we are concerned with predicting unknown values of multiple coupled tensors, our goal is to leverage the relational information from all the views to help predict the unlabeled instances, as well as to use the complementary among different views to improve the performance. Specifically, we are interested in finding a predictive function $f : \mathfrak{X}^{(1)} \times \mathfrak{X}^{(2)} \times \dots \times \mathfrak{X}^{(V)} \rightarrow \mathfrak{Y}$ that minimizes the expected loss, where $\mathfrak{X}^{(v)}$, $v \in [1 : V]$ is the input space in the v -th view and \mathfrak{Y} is the output space.

3 METHODOLOGY

In this section, we first discuss how to design the predictive models for learning from multiple coupled tensors. We then derive structural factorization machines (SFM) that can learn the common latent spaces shared in multi-view coupled tensors and automatically adjust the importance of each view in the predictive model.

3.1 Predictive Models

Without loss of generality, we take two views as an example to introduce our basic design of the predictive models. Specifically, we consider coupled analysis of a third-order tensor and a matrix with one mode in common. Given an input instance $(\{\tilde{\mathcal{X}}^{(1)}, \tilde{\mathcal{X}}^{(2)}\}, y)$, where $\tilde{\mathcal{X}}^{(1)} = \tilde{\mathbf{x}}^{(1)} \circ \tilde{\mathbf{x}}^{(2)} \circ \tilde{\mathbf{x}}^{(3)} \in \mathbb{R}^{(1+I) \times (1+J) \times (1+K)}$ and $\tilde{\mathcal{X}}^{(2)} = \tilde{\mathbf{x}}^{(3)} \circ \tilde{\mathbf{x}}^{(4)} \in \mathbb{R}^{(1+K) \times (1+L)}$. An intuitive solution is to build the following multiple linear model:

$$f(\{\tilde{\mathcal{X}}^{(1)}, \tilde{\mathcal{X}}^{(2)}\}) = \langle \tilde{\mathcal{W}}^{(1)}, \tilde{\mathcal{X}}^{(1)} \rangle + \langle \tilde{\mathcal{W}}^{(2)}, \tilde{\mathcal{X}}^{(2)} \rangle \quad (5)$$

where $\tilde{\mathcal{W}}^{(1)} \in \mathbb{R}^{(1+I) \times (1+J) \times (1+K)}$ and $\tilde{\mathcal{W}}^{(2)} \in \mathbb{R}^{(1+K) \times (1+L)}$ are the weights for each view to be learned.

However, in this case it does not take into account the relations and differences between two views. In order to incorporate the relations between two views and also discriminate the importance

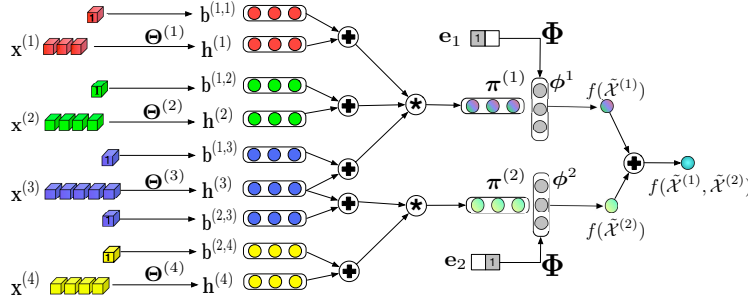


Figure 2: Example of the computational procedure in a structural factorization machine, given the input $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$.

of each view, we introduce an indicator vector $\mathbf{e}_v \in \mathbb{R}^V$ for each view v as

$$\mathbf{e}_v = [\underbrace{0, \dots, 0}_{v-1}, 1, 0, \dots, 0]^T,$$

and transform the predictive model in Eq. (5) into

$$f(\{\tilde{X}^{(1)}, \tilde{X}^{(2)}\}) = \langle \hat{\mathcal{W}}^{(1)}, \tilde{X}^{(1)} \circ \mathbf{e}_1 \rangle + \langle \hat{\mathcal{W}}^{(2)}, \tilde{X}^{(2)} \circ \mathbf{e}_2 \rangle, \quad (6)$$

where $\hat{\mathcal{W}}^{(1)} \in \mathbb{R}^{(1+I) \times (1+J) \times (1+K) \times 2}$ and $\hat{\mathcal{W}}^{(2)} \in \mathbb{R}^{(1+K) \times (1+L) \times 2}$.

Directly learning the weight tensors $\hat{\mathcal{W}}$ s leads to two drawbacks. First, the weight parameters are learned independently for different modes and different views. When the feature interactions rarely (or even never) appear during training, it is unlikely to learn the associated parameters appropriately. Second, the number of parameters in Eq. (6) is exponential to the number of features, which can make the model prone to overfitting and ineffective on sparse data. Here, we assume that each weight tensor has a low-rank approximation, and $\hat{\mathcal{W}}^{(1)}$ and $\hat{\mathcal{W}}^{(2)}$ can be decomposed by CP factorization as

$$\begin{aligned} \hat{\mathcal{W}}^{(1)} &= \llbracket \hat{\boldsymbol{\theta}}^{(1,1)}, \hat{\boldsymbol{\theta}}^{(1,2)}, \hat{\boldsymbol{\theta}}^{(1,3)}, \Phi \rrbracket \\ &= \llbracket [\mathbf{b}^{(1,1)}; \boldsymbol{\theta}^{(1)}], [\mathbf{b}^{(1,2)}; \boldsymbol{\theta}^{(2)}], [\mathbf{b}^{(1,3)}; \boldsymbol{\theta}^{(3)}], \Phi \rrbracket, \end{aligned}$$

and

$$\hat{\mathcal{W}}^{(2)} = \llbracket \hat{\boldsymbol{\theta}}^{(2,3)}, \hat{\boldsymbol{\theta}}^{(2,4)}, \Phi \rrbracket = \llbracket [\mathbf{b}^{(2,3)}; \boldsymbol{\theta}^{(3)}], [\mathbf{b}^{(2,4)}; \boldsymbol{\theta}^{(4)}], \Phi \rrbracket,$$

where $\boldsymbol{\theta}^{(m)} \in \mathbb{R}^{I_m \times R}$ is the factor matrix for the features in the m -th mode. It is worth noting that $\boldsymbol{\theta}^{(3)}$ is shared in the two views. $\Phi \in \mathbb{R}^{2 \times R}$ is the factor matrix for the view indicator, and $\mathbf{b}^{(v,m)} \in \mathbb{R}^{1 \times R}$, which is always associated with the constant one in $\tilde{\mathbf{x}}^{(m)} = [1; \mathbf{x}^{(m)}]$, represents the bias factors of the m -th mode in the v -th view. Through $\mathbf{b}^{(v,m)}$, the lower-order interactions (the interactions excluding the features from the m -th mode) in the v -th view are explored in the predictive function.

Then we can transform Eq. (6) into

$$\begin{aligned} & \langle \hat{\mathcal{W}}^{(1)}, \tilde{X}^{(1)} \circ \mathbf{e}_1 \rangle + \langle \hat{\mathcal{W}}^{(2)}, \tilde{X}^{(2)} \circ \mathbf{e}_2 \rangle \\ &= \sum_{r=1}^R \langle \hat{\theta}_r^{(1,1)} \circ \hat{\theta}_r^{(1,2)} \circ \hat{\theta}_r^{(1,3)} \circ \phi_r, \tilde{\mathbf{x}}^{(1)} \circ \tilde{\mathbf{x}}^{(2)} \circ \tilde{\mathbf{x}}^{(3)} \circ \mathbf{e}_1 \rangle \\ & \quad + \sum_{r=1}^R \langle \hat{\theta}_r^{(2,3)} \circ \hat{\theta}_r^{(2,4)} \circ \phi_r, \tilde{\mathbf{x}}^{(3)} \circ \tilde{\mathbf{x}}^{(4)} \circ \mathbf{e}_2 \rangle \\ &= \phi^1 \left(\prod_{m=1}^3 * (\tilde{\mathbf{x}}^{(m)T} \hat{\boldsymbol{\theta}}^{(1,m)}) \right)^T + \phi^2 \left(\prod_{m=3}^4 * (\tilde{\mathbf{x}}^{(m)T} \hat{\boldsymbol{\theta}}^{(2,m)}) \right)^T \\ &= \phi^1 \left(\prod_{m=1}^3 * (\mathbf{x}^{(m)T} \boldsymbol{\theta}^{(m)} + \mathbf{b}^{(1,m)}) \right)^T + \phi^2 \left(\prod_{m=3}^4 * (\mathbf{x}^{(m)T} \boldsymbol{\theta}^{(m)} + \mathbf{b}^{(2,m)}) \right)^T \end{aligned} \quad (7)$$

where $*$ is the Hadamard (elementwise) product and $\phi^v \in \mathbb{R}^{1 \times R}$ is the v -th row of the factor matrix Φ .

For convenience, we let $\mathbf{h}^{(m)} = \boldsymbol{\theta}^{(m)T} \mathbf{x}^{(m)}$, $S_M(v)$ denote the set of modes in the v -th views, $\boldsymbol{\pi}^{(v)} = \prod_{m \in S_M(v)} * (\mathbf{h}^{(m)} + \mathbf{b}^{(v,m)T})$, and $\boldsymbol{\pi}^{(v,-m)} = \prod_{m' \in S_M(v), m' \neq m} * (\mathbf{h}^{(m')} + \mathbf{b}^{(v,m')T})$. It is worth noting that $\mathbf{h}^{(m)}$ can be regarded as the latent representation of the feature $\mathbf{x}^{(m)}$ in m -th mode, and $\boldsymbol{\pi}^{(v)}$ can be regarded as the joint representation of all the modes in the v -th view, which can be easily computed through the Hadamard product. Moreover, the contribution of the latent factors in $\boldsymbol{\pi}^{(v)}$ is automatically adjusted by the weight vector ϕ^v . An graphical illustration of the computational procedure in Eq. (7) is shown in Fig. 2.

The predictive model for the general cases is given as follows

$$\begin{aligned} f(\{\tilde{X}^{(v)}\}) &= \sum_{v=1}^V \langle \hat{\mathcal{W}}^{(v)}, \tilde{X}^{(v)} \circ \mathbf{e}_v \rangle \\ &= \sum_{v=1}^V \phi^v \prod_{m \in S_M(v)} * (\mathbf{x}^{(m)T} \boldsymbol{\theta}^{(m)} + \mathbf{b}^{(v,m)T})^T \quad (8) \\ &= \sum_{v=1}^V \phi^v \prod_{m \in S_M(v)} * (\mathbf{h}^{(m)} + \mathbf{b}^{(v,m)T}) \end{aligned}$$

We name this model as structural factorization machines (SFMs). Clearly, the parameters are jointly factorized, which benefits parameter estimation under sparsity since dependencies exist when

the interactions share the same features. Therefore, the model parameters can be effectively learned without direct observations of such interactions especially in highly sparse data. More importantly, after factorizing the weight tensor \mathcal{W} s, there is no need to construct the input tensor physically. Furthermore, the model complexity is linear in the number of original features. In particular, the model complexity is $O(R(V + I + \sum_v M_v))$, where M_v is the number of modes in the v -th view.

3.2 Learning Structural Factorization Machines

Following the traditional supervised learning framework, we propose to learn the model parameters by minimizing the following regularized empirical risk:

$$\mathcal{R} = \frac{1}{N} \sum_{n=1}^N \ell(f(\{X_n^{(v)}\}), y_n) + \lambda \Omega(\Phi, \{\Theta^{(m)}\}, \{\mathbf{b}^{(v,m)}\}) \quad (9)$$

where ℓ is a prescribed loss function, Ω is the regularizer encoding the prior knowledge of $\{\Theta^{(m)}\}$ and Φ , and $\lambda > 0$ is the regularization parameter that controls the trade-off between the empirical loss and the prior knowledge.

The partial derivative of \mathcal{R} w.r.t. $\Theta^{(m)}$ is given by

$$\frac{\partial \mathcal{R}}{\partial \Theta^{(m)}} = \frac{\partial \mathcal{L}}{\partial f} \frac{\partial f}{\partial \Theta^{(m)}} + \lambda \frac{\partial \Omega_\lambda(\Theta^{(m)})}{\partial \Theta^{(m)}} \quad (10)$$

where $\frac{\partial \mathcal{L}}{\partial f} = \frac{1}{N} \left[\frac{\partial \ell}{\partial f}, \dots, \frac{\partial \ell}{\partial f} \right]^T \in \mathbb{R}^N$.

For convenience, we let $S_V(m)$ denote the set of views that contains the m -th mode, $\mathbf{X}^{(m)} = [\mathbf{x}_1^{(m)}, \dots, \mathbf{x}_N^{(m)}]$, $\Pi^{(v)} = [\pi_1^{(v)}, \dots, \pi_N^{(v)}]^T$ and $\Pi^{(v,-m)} = [\pi_1^{(v,-m)}, \dots, \pi_N^{(v,-m)}]^T$. We then have that

$$\frac{\partial \mathcal{L}}{\partial f} \frac{\partial f}{\partial \Theta^{(m)}} = \mathbf{X}^{(m)} \left(\sum_{v \in S_V(m)} \left(\left(\frac{\partial \mathcal{L}}{\partial f} \phi^v \right) * \Pi^{(v,-m)} \right) \right) \quad (11)$$

Similarly, the partial derivative of \mathcal{R} w.r.t. $\mathbf{b}^{(v,m)}$ is given by

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial \mathbf{b}^{(v,m)}} &= \frac{\partial \mathcal{L}}{\partial f} \frac{\partial f}{\partial \mathbf{b}^{(v,m)}} + \lambda \frac{\partial \Omega_\lambda(\mathbf{b}^{(v,m)})}{\partial \mathbf{b}^{(v,m)}} \\ &= \mathbf{1}^T \left(\left(\frac{\partial \mathcal{L}}{\partial f} \phi^v \right) * \Pi^{(v,-m)} \right) + \lambda \frac{\partial \Omega_\lambda(\mathbf{b}^{(v,m)})}{\partial \mathbf{b}^{(v,m)}} \end{aligned} \quad (12)$$

The partial derivative of \mathcal{R} w.r.t. Φ is given by

$$\frac{\partial \mathcal{R}}{\partial \Phi} = \left[\left(\frac{\partial \mathcal{L}}{\partial f} \right)^T \Pi^{(1)}; \dots; \left(\frac{\partial \mathcal{L}}{\partial f} \right)^T \Pi^{(V)} \right] + \lambda \frac{\partial \Omega_\lambda(\Phi)}{\partial \Phi} \quad (13)$$

Finally, the gradient of \mathcal{R} can be formed by vectorizing the partial derivatives with respect to each factor matrix and concatenating them all, i.e.,

$$\nabla \mathcal{R} = \begin{bmatrix} \text{vec}\left(\frac{\partial \mathcal{R}}{\partial \Theta^{(1)}}\right) \\ \vdots \\ \text{vec}\left(\frac{\partial \mathcal{R}}{\partial \Theta^{(M)}}\right) \\ \text{vec}\left(\frac{\partial \mathcal{R}}{\partial \mathbf{b}^{(1,1)}}\right) \\ \vdots \\ \text{vec}\left(\frac{\partial \mathcal{R}}{\partial \mathbf{b}^{(V,M)}}\right) \\ \text{vec}\left(\frac{\partial \mathcal{R}}{\partial \Phi}\right) \end{bmatrix} \quad (14)$$

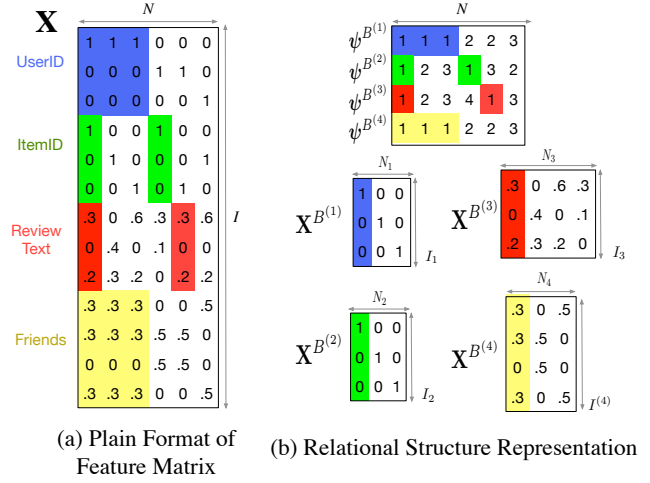


Figure 3: (a) Feature vectors of the same entity repeatedly appear in the plain formatted feature matrix \mathbf{X} . (b) Repeating patterns in \mathbf{X} can be formalized by the relational structure \mathbf{B} of each mode. For example, the forth column of the feature matrix \mathbf{X} can be represented as $\mathbf{x}_4 = [\mathbf{x}_{\psi^{(1)}(4)}^{(1)}; \mathbf{x}_{\psi^{(2)}(4)}^{(2)}; \mathbf{x}_{\psi^{(3)}(4)}^{(3)}; \mathbf{x}_{\psi^{(4)}(4)}^{(4)}] = [\mathbf{x}_2^{B(1)}; \mathbf{x}_1^{B(2)}; \mathbf{x}_4^{B(3)}; \mathbf{x}_2^{B(4)}]$.

Once we have the function, \mathcal{R} and gradient, $\nabla \mathcal{R}$, values, we can use any gradient-based optimization algorithm to compute the factor matrices. For the results presented in this paper, we use the Adam optimization algorithm [11] for parameter updates.

3.3 Efficient Computing with Relational Structures

In relational domains, we can often observe that feature vectors of the same entity repeatedly appear in the plain formatted feature matrix \mathbf{X} , where $\mathbf{X} = [\mathbf{X}^{(1)}; \dots; \mathbf{X}^{(M)}] \in \mathbb{R}^{I \times N}$ and $\mathbf{X}^{(m)} \in \mathbb{R}^{I_m \times N}$ is the feature matrix in the m -th mode. Consider Fig. 3(a) as an example, where the parts highlighted in yellow in the forth mode (which represents the friends of the user) are repeatedly appear in the first three columns. Clearly, these repeating patterns stem from the relational structure of the same entity.

In the following, we show how the proposed SFM method can make use of relational structure of each mode, such that the learning and prediction can be scaled to predictor variables generated from relational data involving relations of high cardinality. We adopt the idea from [23] to avoid redundant computing on repeating patterns over a set of feature vectors.

Let $\mathcal{B} = \{(\mathbf{X}^{B(m)}, \psi^{B(m)})\}_{m=1}^M$ be the set of relational structures, where $\mathbf{X}^{B(m)} \in \mathbb{R}^{I_m \times N_m}$ denotes the relational matrix of m -th mode, $\psi^{B(m)} : \{1, \dots, N\} \rightarrow \{1, \dots, N_m\}$ denotes the mapping from columns in the feature matrix \mathbf{X} to columns within $\mathbf{X}^{B(m)}$. To shorten notation, the index B is dropped from the mapping ψ^B whenever it is clear which block the mapping belongs to. From \mathcal{B} , one can reconstruct \mathbf{X} by concatenating the corresponding columns

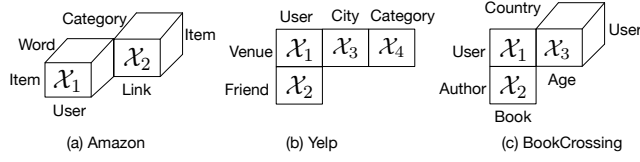


Figure 4: Schema of the structural views in each dataset.

of the relational matrices using the mappings. For instance, the feature vector \mathbf{x}_n of the n -th case in the plain feature matrix \mathbf{X} is represented as $\mathbf{x}_n = [\mathbf{x}_{\psi(n)}^{(1)}; \dots; \mathbf{x}_{\psi(n)}^{(M)}]$. Fig. 3(b) shows an example how the feature matrix can be represented in relational structures. Let $N_z(\mathbf{A})$ denote the number of non-zeros in a matrix \mathbf{A} . The space required for using relational structures to represent the input data is $|\mathcal{B}| = NM + \sum_m N_z(\mathbf{X}^{B(m)})$, which is much smaller than $N_z(\mathbf{X})$ if there are repeating patterns in the feature matrix \mathbf{X} .

Now we can rewrite the predictive model in Eq. (8) as follows

$$f(\{\mathcal{X}_n^{(v)}\}) = \sum_{v=1}^V \phi^v \prod_{m \in S_M(v)} * (\mathbf{h}_{\psi(n)}^{B(m)} + \mathbf{b}^{(v,m)^T}), \quad (15)$$

with the caches $\mathbf{H}^{B(m)} = [\mathbf{h}_1^{B(m)}, \dots, \mathbf{h}_{N_m}^{B(m)}]$ for each mode, where $\mathbf{h}_j^{B(m)} = \Theta^{(m)^T} \mathbf{x}_j^{B(m)}$, $\forall j \in [1 : N_m]$.

This directly shows how N samples can be efficiently predicted:

(i) compute $\mathbf{H}^{B(m)}$ in $O(RN_z(\mathbf{X}^{B(m)}))$ for each mode, (ii) compute N predictions with Eq. (15) using caches in $O(RN(V + \sum_v M_v))$. With the help of relational structures, SFMs can learn the same parameters and make the same predictions but with a much lower runtime complexity.

4 EXPERIMENTS

4.1 Datasets

To evaluate the ability and applicability of the proposed SFMs, we include a spectrum of large datasets from different domains. The statistics for each dataset is summarized in Table 2, the schema of the structural views in each dataset is presented in Fig. 4, and the details are as follows:

Amazon²: The first group of datasets are from Amazon.com recently introduced by [20]. This is among the largest datasets available that includes review texts and metadata of items. Each top-level category of products on Amazon.com has been constructed as an independent dataset in [20]. In this paper, we take a variety of large categories as listed in Tabel 2.

Each sample in these datasets has five modes, *i.e.*, users, items, review texts, categories, and linkage. The user mode and item mode are represented by one-hot encoding. The ℓ_2 -normalized TF-IDF vector representation of review text³ of the item given by the user is used as the text mode. The category mode and linkage mode consists of all the categories and all the co-purchasing items of the item, which might be from other categories. The last two modes are ℓ_1 -normalized.

²<http://jmcauley.ucsd.edu/data/amazon/>

³Stemming, lemmatization, removing stop-words and words with frequency less than 100 times, etc., are handled beforehand.

Yelp⁴: It is a large-scale dataset consisting of venue reviews. Each sample in this dataset contains five modes, *i.e.*, users, venues, friends, categories and cities. The user mode and venue mode are represented by one-hot encoding. The friend mode consists of the friends' ids of users. The category mode and city mode consists of all the categories and the city of the venue. The last three modes are ℓ_1 -normalized.

BookCrossing (BX)⁵: It is a book review dataset collected from the Book-Crossing community. Each sample in this dataset contains five modes, *i.e.*, users, books, countries, ages and authors. The ages are split in eight bins as in [7]. The country mode and age mode consist of the corresponding meta information of the user. The author modes represents the authors of the book. All the modes are represented by one-hot encoding.

The values of samples range within [1:5] in Amazon and Yelp datasets, and range within [1:10] in BX dataset.

4.2 Comparison Methods

In order to demonstrate the effectiveness of the proposed SFMs, we compare a series of state-of-the-art methods.

Matrix Factorization (MF) is used to validate that meta information is helpful for improving prediction performance. We use the LIBMF implementation [6] for comparison in the experiment.

Factorization Machine (FM) [22] is the state-of-the-art method in recommender systems. We compare with its higher-order extension [2] with up to second-order, and third-order feature interactions, and denote them as FM-2 and FM-3.

Polynomial Network (PolyNet) [16] is a recently proposed method that utilizes polynomial kernel on all features. We compare the augmented PolyNet (which adds a constant one to the feature vector [3]) with up to the second-order, and third-order kernel and denote them as PolyNet-2 and PolyNet-3.

Multi-View Machine (MVM) [5] is a tensor factorization based method that explores the latent representation embedded in the full-order interactions among all the modes.

Structural Factorization Machine (SFM) is the proposed model that learns the common latent spaces shared in structural data.

4.3 Experimental Settings

For each dataset, we randomly split 50%, 10%, and 40% of labeled samples as training set, validation set, and testing set, respectively. Validation sets are used for hyper-parameter tuning for each model. Each of the validation and testing sets does not overlap with any other set so as to ensure the sanity of the experiment. For simplicity and fair comparison, in all the comparison methods, the dimension of latent factors $R = 20$ and the maximum number of epochs is set as 400, and Forbenius norm regularizers are used to avoid overfitting. The regularization hyper-parameter is tuned from $\{10^{-5}, 10^{-4}, \dots, 10^0\}$.

All the methods except MF are implemented in TensorFlow, and the parameters are initialized using scaling variance initializer [8], and Adam optimizer [11]. We tune the scaling factor of initializer σ from $\{1, 2, 5, 10, 100\}$ and the learning rate η from $\{0.01, 0.1, 1\}$

⁴<https://www.yelp.com/dataset-challenge>

⁵<http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

Table 2: The statistics for each dataset. $N_z(X)$ and $N_z(\mathcal{B})$ are the number of non-zeros in plain formatted feature matrix and in relational structures, respectively. Game: Video Games, Cloth: Clothing, Shoes and Jewelry, Sport: Sports and Outdoors, Health: Health and Personal Care, Home: Home and Kitchen, Elec: Electronics.

Dataset	#Samples	Mode					Density	$N_z(X)$	$N_z(\mathcal{B})$
Amazon		#Users	#Items	#Words	#Categories	#Links			
Game	231,780	24,303	10,672	7,500	193	17,974	0.089%	32.9M	15.2M
Cloth	278,677	39,387	23,033	3,493	1,175	107,139	0.031%	25.6M	7.3M
Sport	296,337	35,598	18,357	5,202	1,432	73,040	0.045%	34.2M	10.2M
Health	346,355	38,609	18,534	5,889	849	80,379	0.048%	33.6M	12.1M
Home	551,682	66,569	28,237	6,455	970	99,090	0.029%	46.8M	19.4M
Elec	1,689,188	192,403	63,001	12,805	967	89,259	0.014%	161.5M	69M
Yelp		#Users	#Venues	#Friends	#Categories	#Cities			
Yelp	1,319,870	88,009	40,520	88,009	892	412	0.037%	70.5M	1.4M
BX		#Users	#Books	#Countries	#Ages	#Authors			
BX	244,848	24,325	45,074	57	8	17,178	0.022%	1.2M	163K

Table 3: MSE comparison on all the datasets. The best results are listed in bold.

Dataset	(a) MF	(b) MVM	(c) FM-2	(d) FM-3	(e) PolyNet-2	(f) PolyNet-3	(g) SFM	Improvement of SFM versus		
								b	min(c,d)	min(e,f)
Game	1.569 ± 0.005	0.753 ± 0.007	0.764 ± 0.006	0.749 ± 0.007	0.749 ± 0.004	0.748 ± 0.006	0.723 ± 0.006	4.06%	3.52%	3.35%
Cloth	1.624 ± 0.009	0.725 ± 0.046	0.678 ± 0.004	0.679 ± 0.004	0.678 ± 0.007	0.680 ± 0.005	0.659 ± 0.013	9.03%	2.82%	2.84%
Sport	1.290 ± 0.004	0.646 ± 0.019	0.638 ± 0.003	0.632 ± 0.007	0.631 ± 0.005	0.632 ± 0.005	0.614 ± 0.011	5.00%	2.91%	2.79%
Health	1.568 ± 0.007	0.807 ± 0.012	0.779 ± 0.004	0.778 ± 0.004	0.779 ± 0.005	0.776 ± 0.005	0.763 ± 0.019	5.47%	2.02%	1.77%
Home	1.591 ± 0.004	0.729 ± 0.067	0.714 ± 0.002	0.714 ± 0.004	0.690 ± 0.003	0.692 ± 0.005	0.678 ± 0.008	6.93%	5.00%	1.72%
Elec	1.756 ± 0.002	0.792 ± 0.042	0.776 ± 0.006	0.749 ± 0.007	0.760 ± 0.004	0.757 ± 0.001	0.747 ± 0.006	5.69%	0.27%	1.33%
Yelp	1.713 ± 0.003	1.2575 ± 0.013	1.277 ± 0.002	1.277 ± 0.002	1.272 ± 0.002	1.272 ± 0.002	1.256 ± 0.010	0.09%	1.58%	1.19%
BX	4.094 ± 0.025	2.844 ± 0.024	2.766 ± 0.012	2.767 ± 0.014	2.654 ± 0.013	2.658 ± 0.013	2.541 ± 0.025	10.66%	8.16%	4.27%
Average on all datasets								5.87%	3.29%	2.41%

using the validation sets. In the experiment, we set $\sigma = 2$ (default setting in TensorFlow) and $\eta = 0.01$ for these methods except MVM. We found that MVM is more sensitive to the configuration, because MVM will element-wisely multiply the latent factors of all the modes which leads to an extremely small value approaching zero. $\sigma = 10$ and $\eta = 0.1$ yielded the best performance for MVM.

To investigate the performance of comparison methods, we adopt mean squared error (MSE) on the test data as the evaluation metrics [19, 27]. The smaller value of the metric indicates the better performance. Each experiment was repeated for 10 times, and the mean and standard deviation of each metric in each data set were reported. All experiments are conducted on a single machine with Intel Xeon 6-Core CPUs of 2.4 GHz and equipped with a Maxwell Titan X GPU. The code has been made available at GitHub⁶.

4.4 Performance Analysis

The experimental results are shown in Table 3. The best method of each dataset is in bold. For clarity, on the right of the tables we show the percentage improvement of the proposed SFM method over a variety of methods. From these results, we can observe that SFM consistently outperforms all the comparison methods. We also make a few comparisons and summarize our findings as follows.

Compared with MF, SFM performs better with an average improvement of nearly 50%. MF usually performs well in practice [15, 22], while in datasets which are extremely sparse, as is shown in our case, MF is unable to learn an accurate representation of users/items. Thus MF under-performs other methods which takes the meta information into consideration.

In both FM and PolyNet methods, the feature vectors from all the modes are concatenated as a single input feature vector. The major difference between these two methods is the choice of kernel applied [2]. The polynomial kernel used in PolyNet considers all monomials (the products of features), i.e., all combinations of features *with* replacement. The ANOVA kernel used in FM considers only monomials composed of distinct features, i.e., feature combinations *without* replacement. Compared with the best results obtained from FM methods and from PolyNet methods, SFM leads to an average improvement of 3.3% and 2.4% in MSE, respectively.

The primary reason behind the results is how the latent factors of each feature are learned. For any factorization based method, the latent factors of a feature are essentially learned from its interactions with other features observed in the data, as can be observed from its update rule. In FM and PolyNet, all the feature interactions are taken into consideration without distinguishing the features from different modes. As a result, important feature interactions (e.g., the interactions between the given user and her friend) would be easily buried in irrelevant feature interactions

⁶<https://github.com/kevin-luc/SFM>

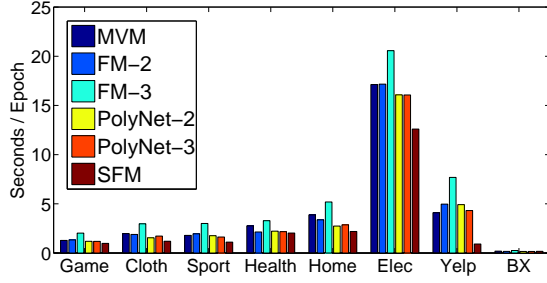


Figure 5: Training Time (Seconds/Epoch) Comparison.

from the same modes (e.g., the interactions between the friends of the same user). Hence, the learned latent factors are less representative in FM and PolyNet, compared with the proposed SFM. Besides, we can find that including higher-order interactions in FM and PolyNet (i.e., FM-3 and PolyNet-3) does not always improve the performance. Instead, it may even degrade the performance, as shown in Cloth, Yelp, and BX datasets. This is probably due to overfitting, as they need to include more parameters to model the interactions in higher orders while the datasets are extremely sparse such that the parameters cannot be properly learned.

Compared to the MVM method, which models the full-order interactions among all the modes, our proposed SFM leads to an average improvement of 5.87%. This is because not all the modes are relevant, and some irrelevant features interactions may introduce unexpected noise to the learning task. The irrelevant information can even be exaggerated after combinations, thereby degrading performance. This suggests that preserving the nature of relational structure is important in building predictive models.

4.5 Computational Cost Analysis

Next, we investigate the computational cost for comparison methods. The training time (seconds per epoch) required for each datasets are shown in Fig. 5. We can easily find that the proposed SFM requires much less computational cost on all the datasets, especially for the Yelp dataset (roughly 11% of computational cost required for training FM-3). The efficiency comes from the use of relational structure representation. As shown in Table 2, the number of non-zeros of the feature matrix $N_z(\mathbf{X})$ is much larger than the number of non-zeros of the relational structure representation $N_z(\mathcal{B})$. The amount of repeating patterns is much higher for the Yelp dataset than for the other dataset, because adding all the friends of a user significantly increases results in large repeating blocks in the plain feature matrix. Standard ML algorithms like the compared methods have typically at best a linear complexity in $N_z(\mathbf{X})$, while using the relational structure representation for SFM have a linear complexity in $N_z(\mathcal{B})$. This experiment substantiates the efficiency of the proposed SFM for large datasets.

4.6 Analysis of the Impact of Data Sparsity

We proceed by further studying the impact of data sparsity on different methods. As can be found in the experimental results,

the improvement of SFM over the traditional collaborative filtering methods (e.g., MF) is significant for datasets that are sparse, mainly because the number of samples is too scarce to model the items and users adequately. We verify this finding by comparing the performance of comparison methods with MF on users with limited training data. Shown in Fig. 6 is the gain of each method compared with MF for users with limited training samples, where G_1 , G_2 , and G_3 are groups of users with [1, 3], [4, 6], and [7, 10] observed samples in the training set. Due to space limit, we only report the results from two Amazon datasets (Sport and Health) while the observations still hold for the rest datasets. It can be seen that the proposed SFM gains the most in group G_1 , in which the users have extremely few training items. The performance gain starts to decrease with the number of training items available for each user. The results indicate that including meta information can be valuable information especially when limited information available.

4.7 Sensitivity analysis

The number of latent factors R is an important hyperparameter for the factorization models. We analyze different values of R and report the results in Fig. 7. The results again demo that SFM consistently outperforms other methods with a various values of R . In contrast to findings in other related models based on latent factors [26] where prediction error can steadily get reduced with larger R , we observe that the performance of each method is rather stable even with the increasing of R . It is reasonable in a general sense, as the expressiveness of the model is enough to describe the information embedded in data. Although larger R renders the model with greater expressiveness, when the available observations regarding the target values are too sparse but the meta information is rich, only a few number of factors is required to fit the data well.

5 CONCLUSIONS

In this paper, we introduce a generic framework for learning structural data from heterogeneous domains, which can explore the high order correlations underlying multi-view structural data. We develop structural factorization machines (SFMs) that learn the common latent spaces shared in the multi-view tensors while automatically adjust the contribution of each view in the predictive model. With the help of relational structure representation, we further provide an efficient approach to avoid unnecessary computation costs on repeating patterns of the multi-view data. It was shown that the proposed SFMs outperform state-of-the-art factorization models on eight large-scale datasets in terms of prediction accuracy and computational cost.

ACKNOWLEDGMENTS

This work is supported in part by NSF through grants IIS-1526499, CNS-1626432, NSFC (61472089, 61503253, 61672357), NSFC-Guangdong Joint Found (U1501254), and the Science Foundation of Guangdong Province (2014A030313556). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

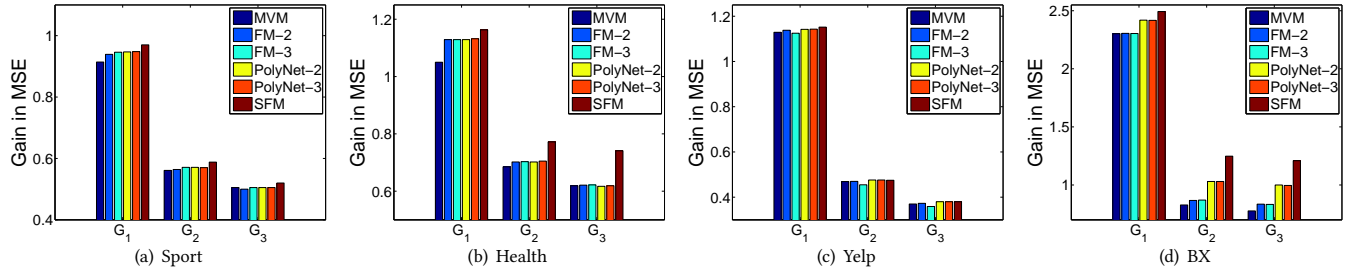


Figure 6: Performance gain in MSE compared with MF for users with limited training samples. G_1 , G_2 , and G_3 are groups of users with [1, 3], [4, 6], and [7, 10] observed samples in the training set, respectively.

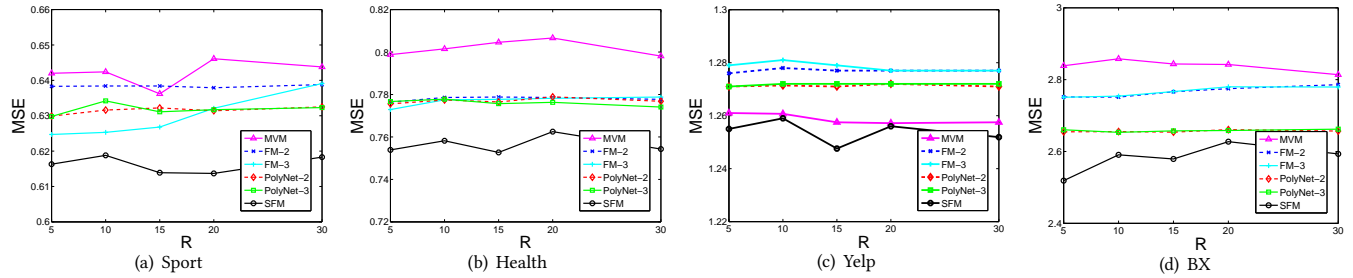


Figure 7: Sensitivity analysis of the latent dimension R .

REFERENCES

- [1] E. Acar, T. G. Kolda, and D. M. Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422*, 2011.
- [2] M. Blondel, A. Fujino, N. Ueda, and M. Ishihata. Higher-order factorization machines. In *Advances in Neural Information Processing Systems*, pages 3351–3359, 2016.
- [3] M. Blondel, M. Ishihata, A. Fujino, and N. Ueda. Polynomial networks and factorization machines: New insights and efficient training algorithms. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016*, pages 850–858, 2016.
- [4] B. Cao, L. He, X. Kong, P. S. Yu, Z. Hao, and A. B. Ragin. Tensor-based multi-view feature selection with applications to brain diseases. In *ICDM*, pages 40–49, 2014.
- [5] B. Cao, H. Zhou, G. Li, and P. S. Yu. Multi-view machines. In *ACM International Conference on Web Search and Data Mining*, pages 427–436, 2016.
- [6] W.-S. Chin, B.-W. Yuan, M.-Y. Yang, Y. Zhuang, Y.-C. Juan, and C.-J. Lin. Libmf: A library for parallel matrix factorization in shared-memory systems. *The Journal of Machine Learning Research*, 17(1):2971–2975, 2016.
- [7] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [9] L. He, X. Kong, S. Y. Philip, A. B. Ragin, Z. Hao, and X. Yang. Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. *matrix*, 3(1):2, 2014.
- [10] L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu. Joint community and structural hole spanner detection via harmonic modularity. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 875–884. ACM, 2016.
- [11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [13] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.
- [14] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):1, 2010.
- [15] G. Ling, M. R. Lyu, and I. King. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 105–112. ACM, 2014.
- [16] R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.
- [17] C.-T. Lu, L. He, W. Shao, B. Cao, and P. S. Yu. Multilinear factorization machines for multi-task multi-view learning. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 701–709. ACM, 2017.
- [18] C.-T. Lu, S. Xie, W. Shao, L. He, and P. S. Yu. Item recommendation for emerging online businesses. In *Proc. 25th Int. Joint Conf. Artificial Intell.*, pages 3797–3803, 2016.
- [19] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.
- [20] J. McAuley, R. Pandey, and J. Leskovec. Inferring networks of substitutable and complementary products. In *SIGKDD*, pages 785–794, 2015.
- [21] S. Rendle. Factorization machines. In *ICDM*, pages 995–1000, 2010.
- [22] S. Rendle. Factorization machines with libFM. *Intelligent Systems and Technology*, 3(3):57, 2012.
- [23] S. Rendle. Scaling factorization machines to relational data. In *Proceedings of the VLDB Endowment*, volume 6, pages 337–348. VLDB Endowment, 2013.
- [24] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90. ACM, 2010.
- [25] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658. ACM, 2008.
- [26] L. Yan, W.-j. Li, G.-R. Xue, and D. Han. Coupled group lasso for web-scale CTR prediction in display advertising. In *ICML*, pages 802–810, 2014.
- [27] L. Zheng, V. Noroozi, and P. S. Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 425–434. ACM, 2017.